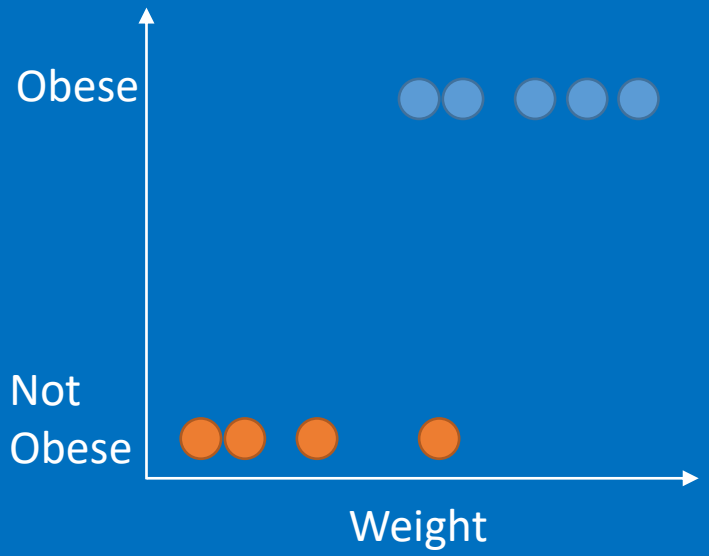
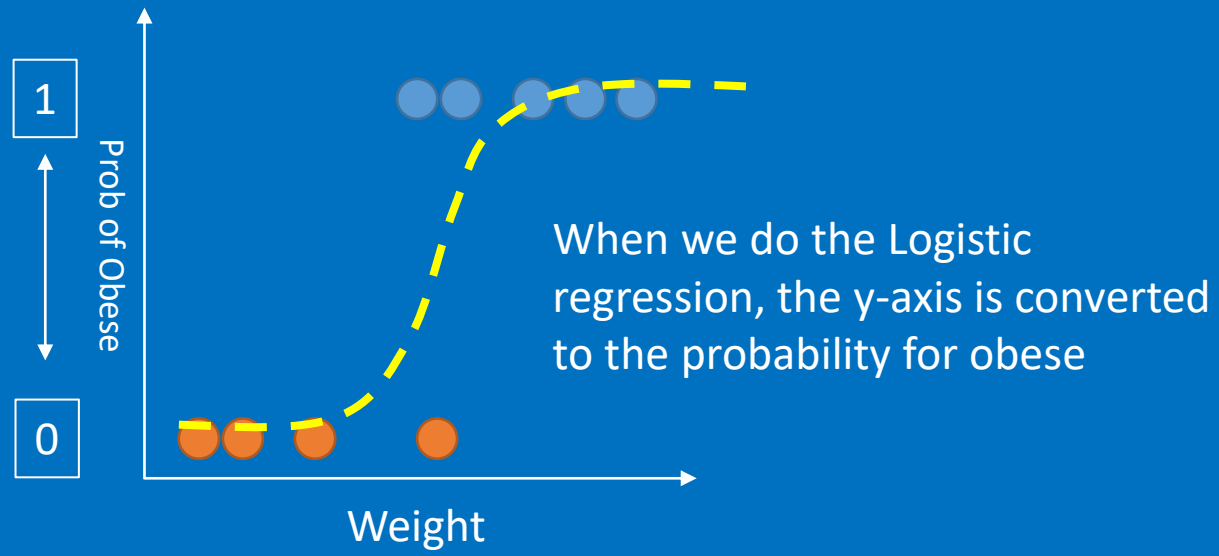
ROC and AUC

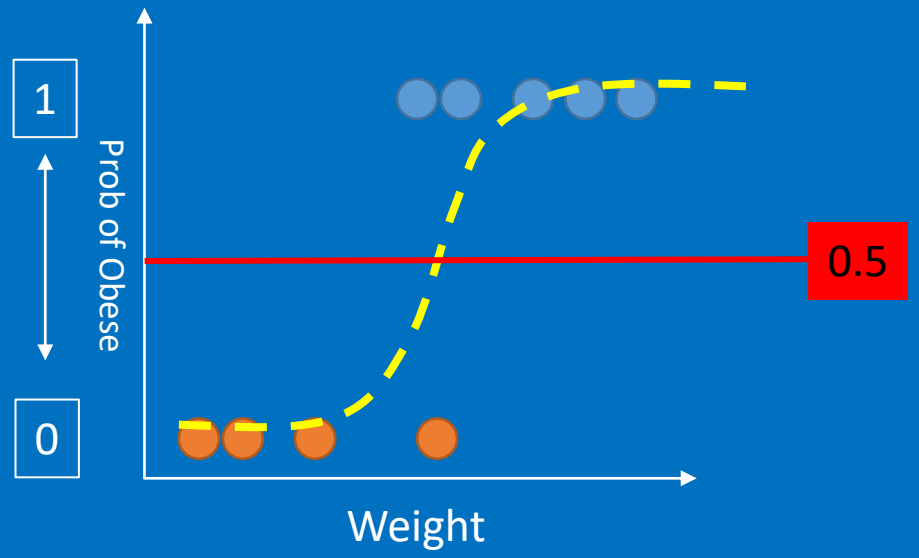


Let’s use logistic regression, and the above data as an example



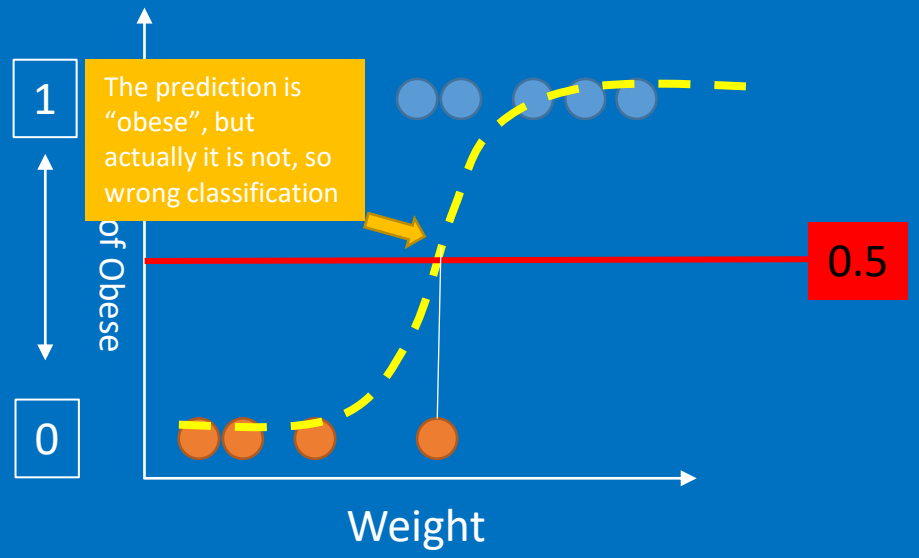
When we do the Logistic regression, the y-axis is converted to the probability for obese

Let’s use logistic regression, and the above data as an example



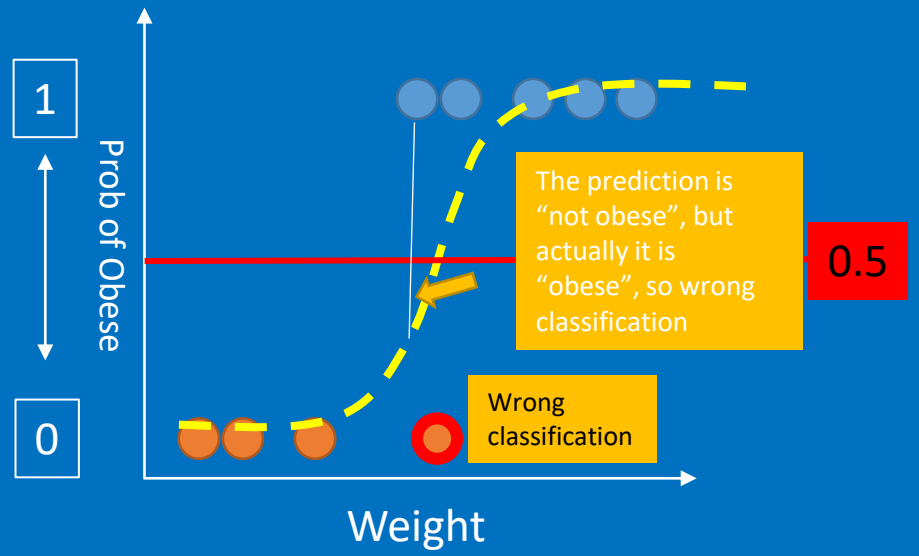
We usually can set 0.5 as the threshold for determining if a man is obese or not

Let’s use logistic regression, and the above data as an example



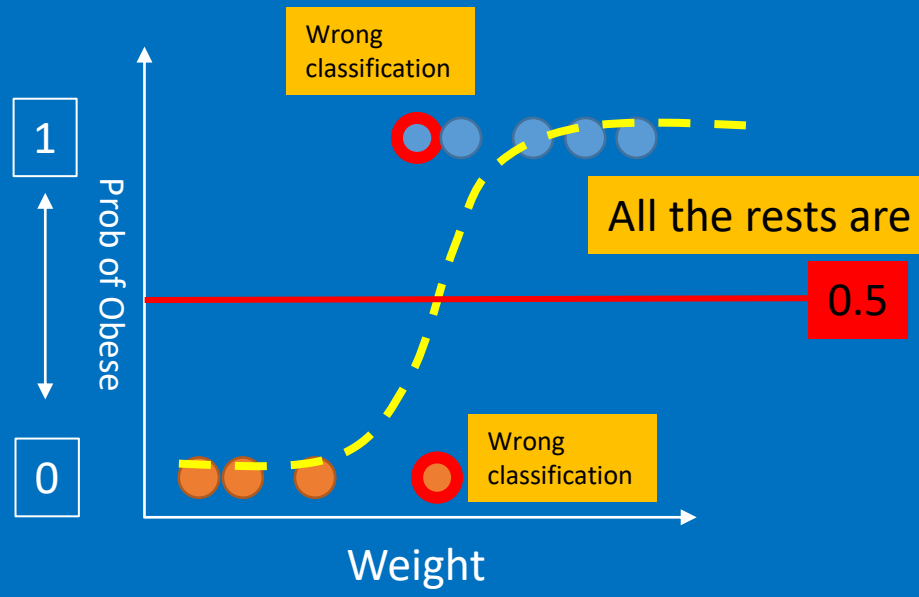
Let's use logistic regression, and the above data as an example

So from the fitted curve, we would know that some points are correctly classified, and some are not



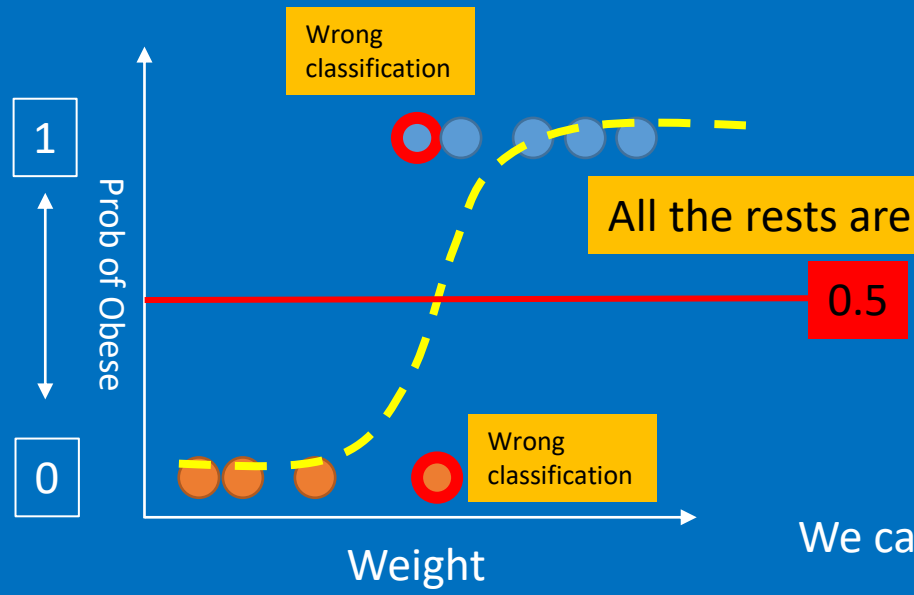
Let's use logistic regression, and the above data as an example

So from the fitted curve, we would know that some points are correctly classified, and some are not



So from the fitted curve, we would know that some points are correctly classified, and some are not

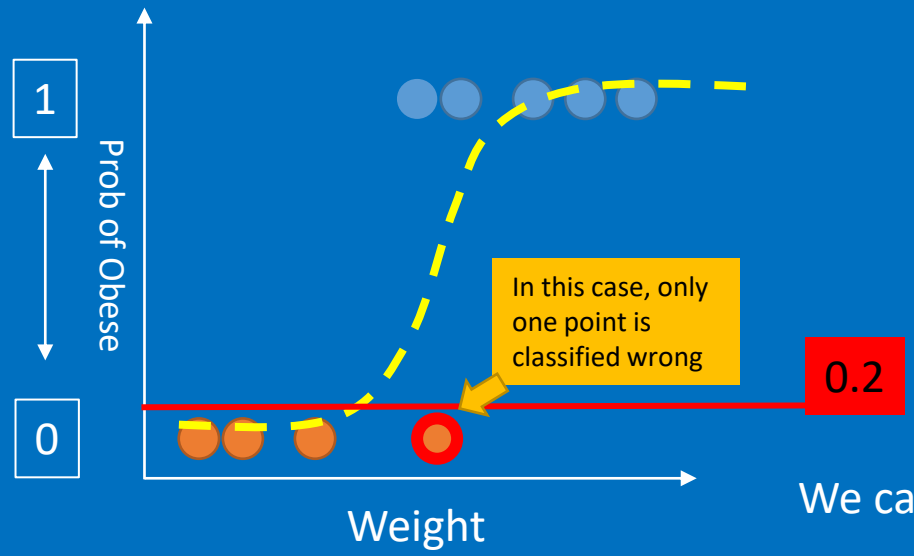
Let’s use logistic regression, and the above data as an example



Let's use logistic regression, and the above data as an example

We can create a confusion matrix to summarize the classification results

Threshold=0.5		**Actual**	
		Is Obese	Not Obese
Prediction	Is Obese	4	1
	Not obese	1	3



So from the fitted curve, we would know that some points are correctly classified, and some are not

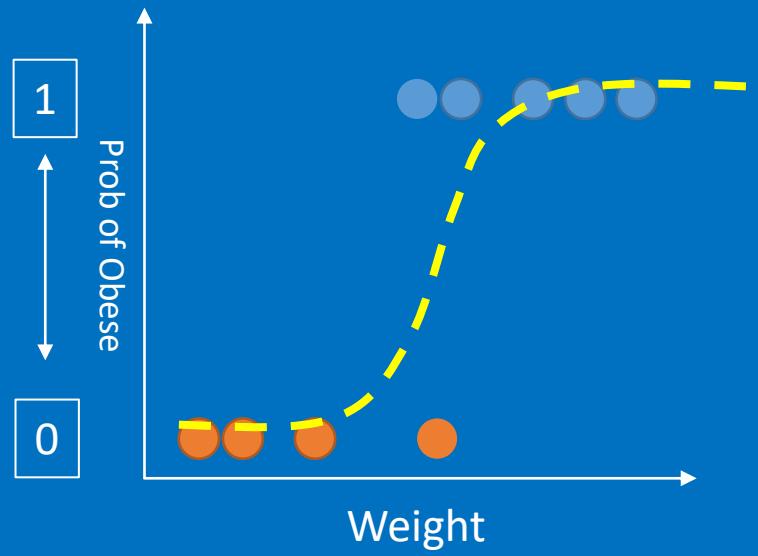
Let's use logistic regression, and the above data as an example

We can create a confusion matrix to summarize the classification results

Threshold=0.5		**Actual**	
		Is Obese	Not Obese
Prediction	Is Obese	4	1
	Not obese	1	3

We can change the threshold to another value, e.g, 0.2, and create another matrix

Threshold=0.2		**Actual**	
		Is Obese	Not Obese
Prediction	Is Obese	5	0
	Not obese	1	3



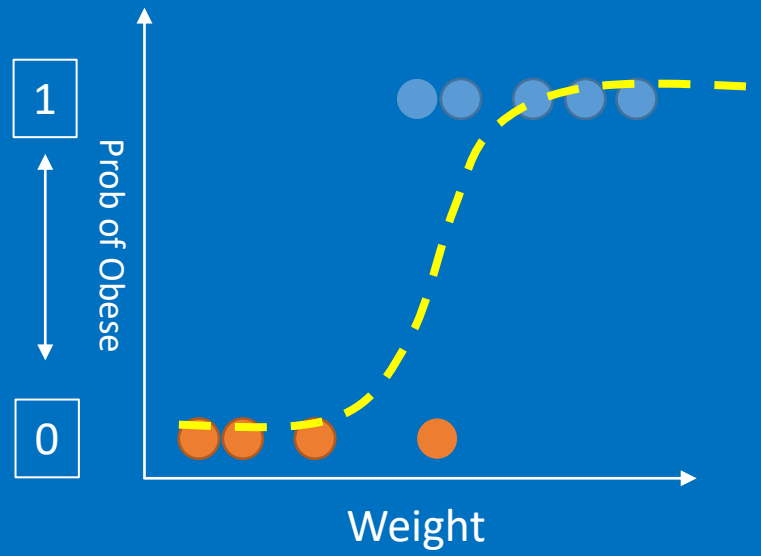
Let’s use logistic regression, and the above data as an example

We can create many confusion matrix for each selected thresholds

Threshold=0.5		Actual	
		Is Obese	Not Obese
Prediction	Is Obese	4	1
	Not obese		

Threshold=0.2		Actual	
		Is Obese	Not Obese
Prediction	Is Obese	5	0
	Not obese	1	3

.....



Let's use logistic regression, and the above data as an example

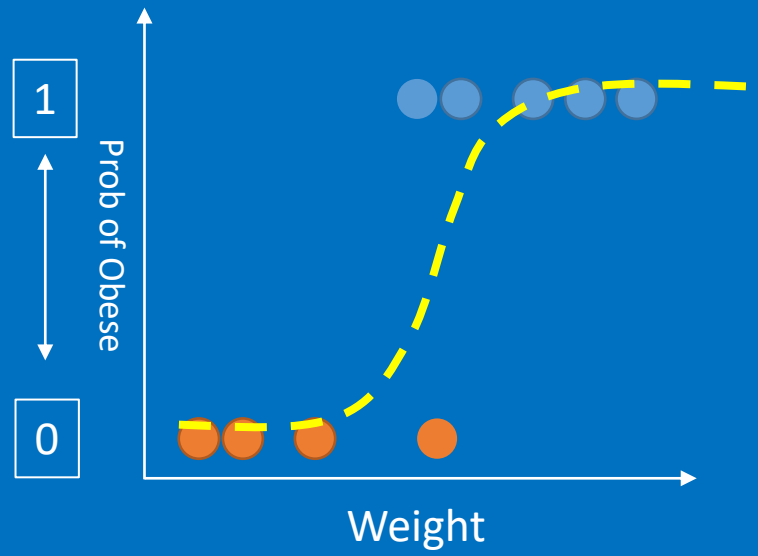
We can create many confusion matrix for each selected thresholds

Threshold=0.5		Actual	
		Is Obese	Not Obese
Prediction	Is Obese	4	1
	Not obese		

Threshold=0.2		Actual	
		Is Obese	Not Obese
Prediction	Is Obese	5	0
	Not obese	1	3

.....

ROC score is designed to simplify the visualization of such huge number of matrices …



Let's use logistic regression, and the above data as an example

We can create many confusion matrix for each selected thresholds

Threshold=0.5		Actual	
		Is Obese	Not Obese
Prediction	Is Obese	4	1
	Not obese		

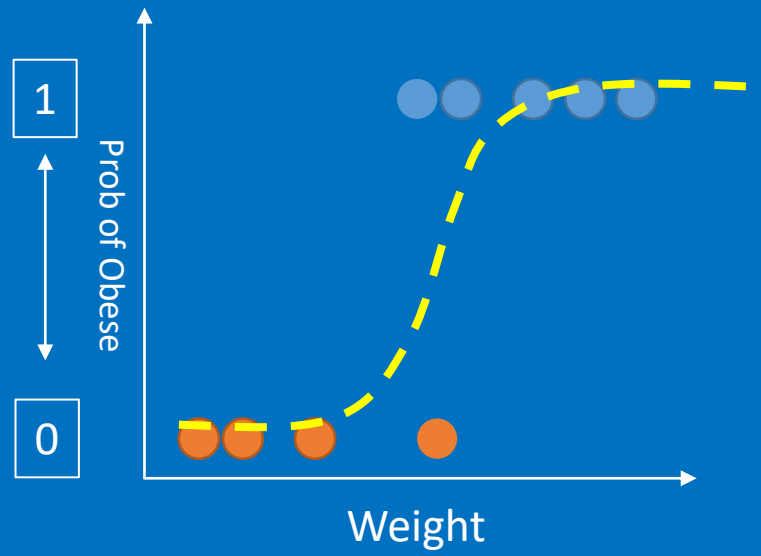
Threshold=0.2		Actual	
		Is Obese	Not Obese
Prediction	Is Obese	5	0
	Not obese	1	3

.....

ROC score is designed to simplify the visualization of such huge number of matrices …

ROC graph has:

- X-axis: True positive rate, or sensitivity
- Y-axis: False positive rate



We can create many confusion matrix for each selected thresholds

Threshold=0.5		Actual	
		Is Obese	Not Obese
Prediction	Is Obese	4	1
	Not obese		

Threshold=0.2		Actual	
		Is Obese	Not Obese
Prediction	Is Obese	5	0
	Not obese	1	3

……

ROC score is designed to simplify the visualization of such huge number of matrices …

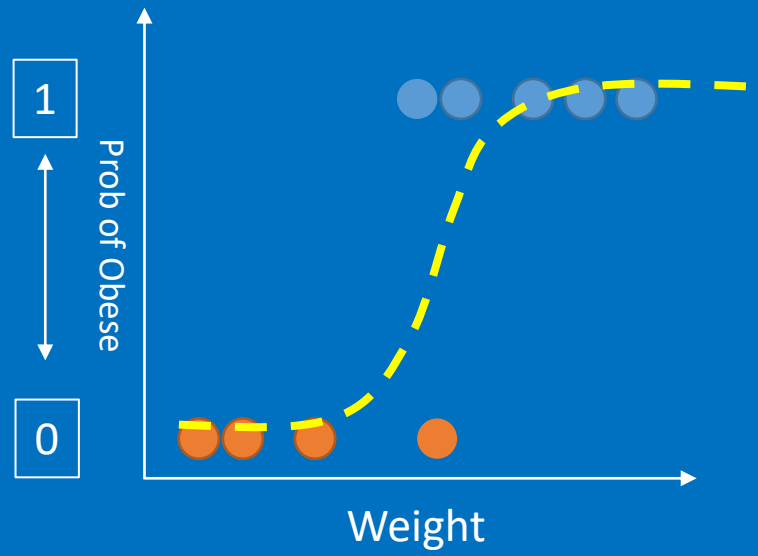
Let's use logistic regression, and the above data as an example

True positive rate can be calculated as

$$True\ positive\ rate = \frac{True\ positives}{True\ positives + False\ negatives}$$

ROC graph has:

- X-axis: True positive rate, or sensitivity
- Y-axis: False positive rate



We can create many confusion matrix for each selected thresholds

Threshold=0.5		Actual	
		Is Obese	Not Obese
Prediction	Is Obese	4	1
	Not obese		

Threshold=0.2		Actual	
		Is Obese	Not Obese
Prediction	Is Obese	5	0
	Not obese	1	3

.....

ROC score is designed to simplify the visualization of such huge number of matrices …

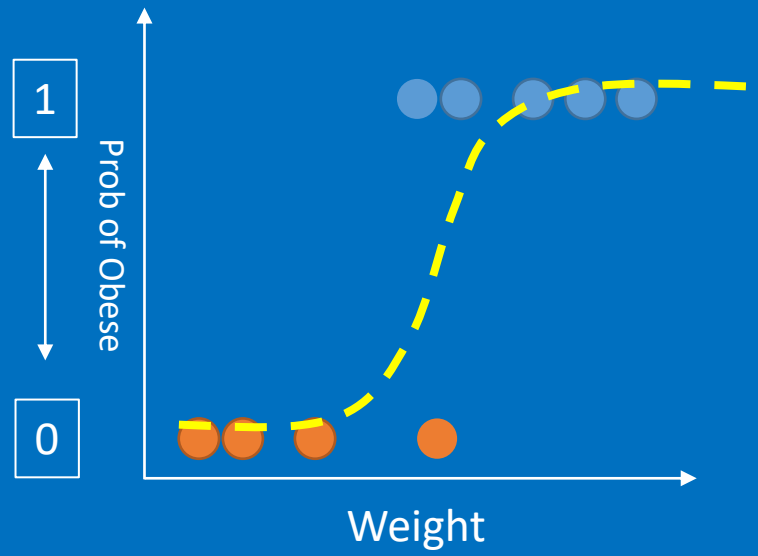
Let's use logistic regression, and the above data as an example

True positive rate can be calculated as

$$\text{True positive rate} = \frac{\text{True positives}}{\text{True positives} + \text{False negatives}}$$

ROC graph has:

- X-axis: True positive rate, or sensitivity
- Y-axis: False positive rate



We can create many confusion matrix for each selected thresholds

Threshold=0.5		Actual	
		Is Obese	Not Obese
Prediction	Is Obese	4	1
	Not obese		

Threshold=0.2		Actual	
		Is Obese	Not Obese
Prediction	Is Obese	5	0
	Not obese	1	3

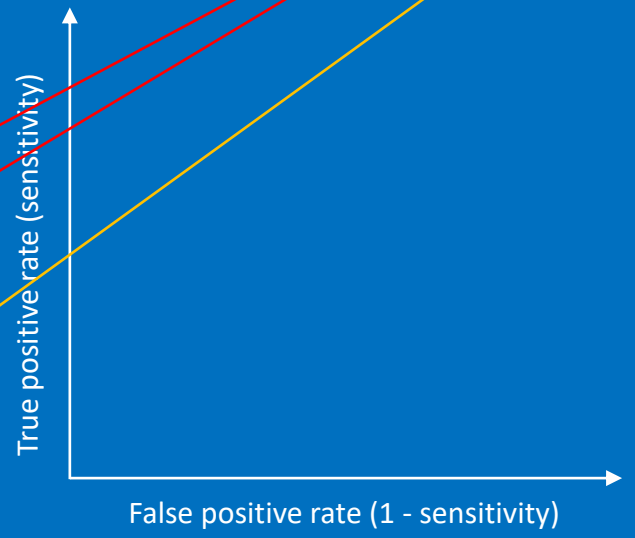
……

ROC score is designed to simplify the visualization of such huge number of matrices …

Let's use logistic regression, and the above data as an example

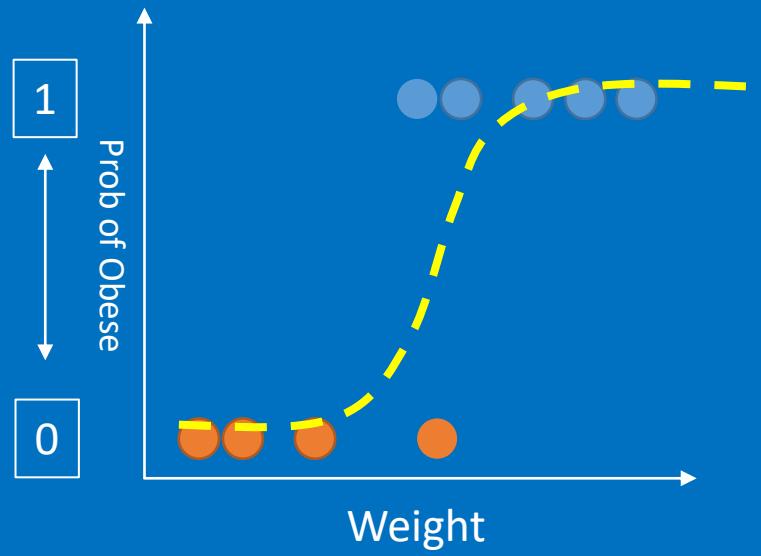
True positive rate can be calculated as

$$\text{True positive rate} = \frac{\text{True positives}}{\text{True positives} + \text{False negatives}}$$



ROC graph has:

- X-axis: True positive rate, or sensitivity
- Y-axis: False positive rate



We can create many confusion matrix for each selected thresholds

Threshold=0.5		Actual	
		Is Obese	Not Obese
Prediction	Is Obese	4	1
	Not obese		

Threshold=0.2		Actual	
		Is Obese	Not Obese
Prediction	Is Obese	5	0
	Not obese	1	3

……

True positive rate tells you what proportion of obese samples that are correctly classified

ROC score is designed to simplify the visualization of such huge number of matrices …

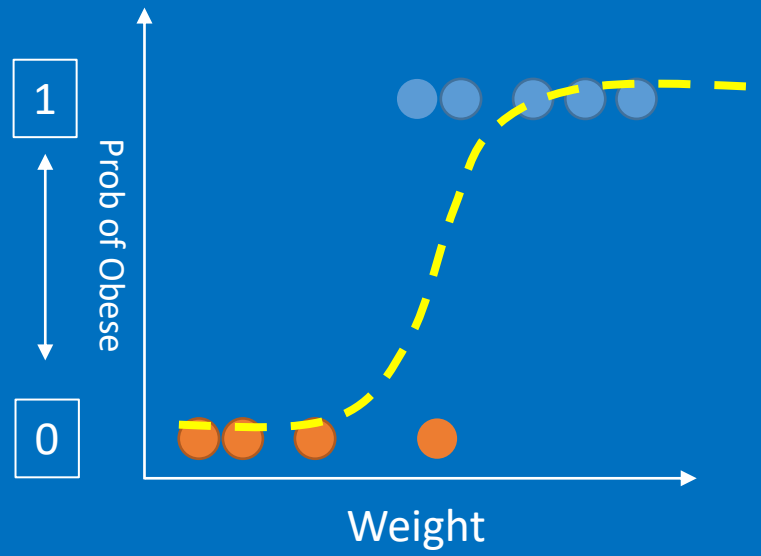
Let's use logistic regression, and the above data as an example

True positive rate can be calculated as

$$\text{True positive rate} = \frac{\text{True positives}}{\text{True positives} + \text{False negatives}}$$

ROC graph has:

- X-axis: True positive rate, or sensitivity
- Y-axis: False positive rate



We can create many confusion matrix for each selected thresholds

Threshold=0.5		Actual	
		Is Obese	Not Obese
Prediction	Is Obese	4	1
	Not obese		

Threshold=0.2		Actual	
		Is Obese	Not Obese
Prediction	Is Obese	5	0
	Not obese	1	3

.....

True positive rate tells you what proportion of obese samples that are correctly classified

ROC score is designed to simplify the visualization of such huge number of matrices …

Let's use logistic regression, and the above data as an example

True positive rate can be calculated as

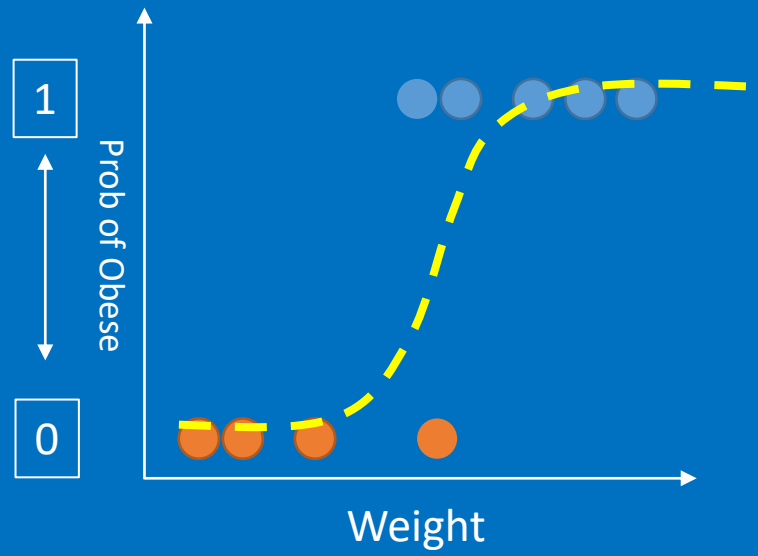
$$True\ positive\ rate = \frac{True\ positives}{True\ positives + False\ negatives}$$

ROC graph has:

- X-axis: True positive rate, or sensitivity
- Y-axis: False positive rate

True False positive rate can be calculated as

$$False\ positive\ rate = \frac{False\ positives}{False\ positives + True\ negatives}$$



Let's use logistic regression, and the above data as an example

True positive rate can be calculated as

$$True\ positive\ rate = \frac{True\ positives}{True\ positives + False\ negatives}$$

We can create many confusion matrix for each selected thresholds

Threshold=0.5		Actual	
		Is Obese	Not Obese
Prediction	Is Obese	4	1
	Not obese		

Threshold=0.2		Actual	
		Is Obese	Not Obese
Prediction	Is Obese	5	0
	Not obese	1	

.....

True positive rate tells you what proportion of obese samples that are correctly classified

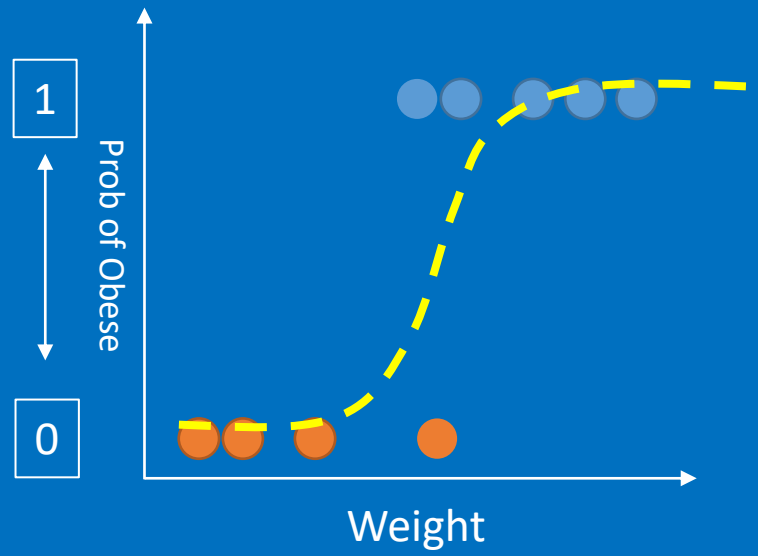
ROC score is designed to simplify the visualization of such huge number of matrices …

ROC graph has:

- X-axis: True positive rate, or sensitivity
- Y-axis: False positive rate

True False positive rate can be calculated as

$$False\ positive\ rate = \frac{False\ positives}{False\ positives + True\ negatives}$$



Let's use logistic regression, and the above data as an example

True positive rate can be calculated as

$$True\ positive\ rate = \frac{True\ positives}{True\ positives + False\ negatives}$$

We can create many confusion matrix for each selected thresholds

Threshold=0.5		Actual	
		Is Obese	Not Obese
Prediction	Is Obese	4	1
	Not obese		

Threshold=0.2		Actual	
		Is Obese	Not Obese
Prediction	Is Obese	5	0
	Not obese	1	3

.....

True positive rate tells you what proportion of obese samples that are correctly classified

False positive rate tells you what proportion of obese samples that are NOT correctly classified

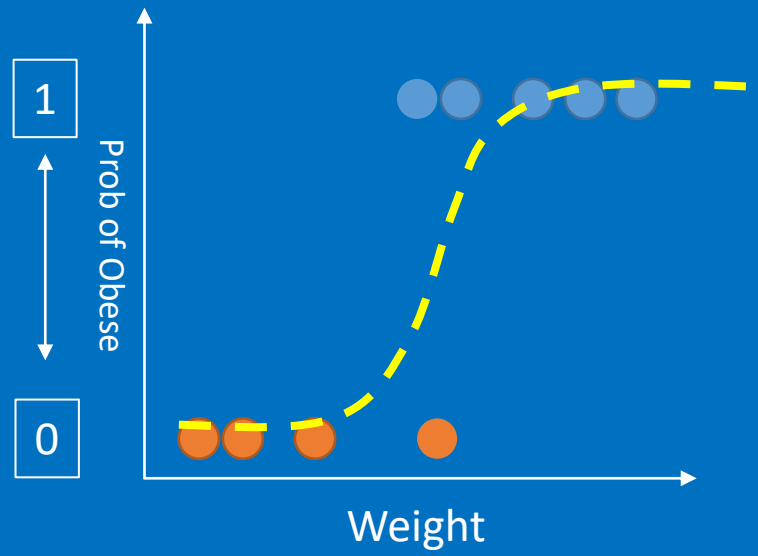
ROC score is designed to simplify the visualization of such huge number of matrices …

ROC graph has:

- X-axis: True positive rate, or sensitivity
- Y-axis: False positive rate

True False positive rate can be calculated as

$$False\ positive\ rate = \frac{False\ positives}{False\ positives + True\ negatives}$$



Let's use logistic regression, and the above data as an example

True positive rate can be calculated as

$$True\ positive\ rate = \frac{True\ positives}{True\ positives + False\ negatives}$$

We can create many confusion matrix for each selected thresholds

Threshold=0.5		Actual	
		Is Obese	Not Obese
Prediction	Is Obese	4	1
	Not obese		

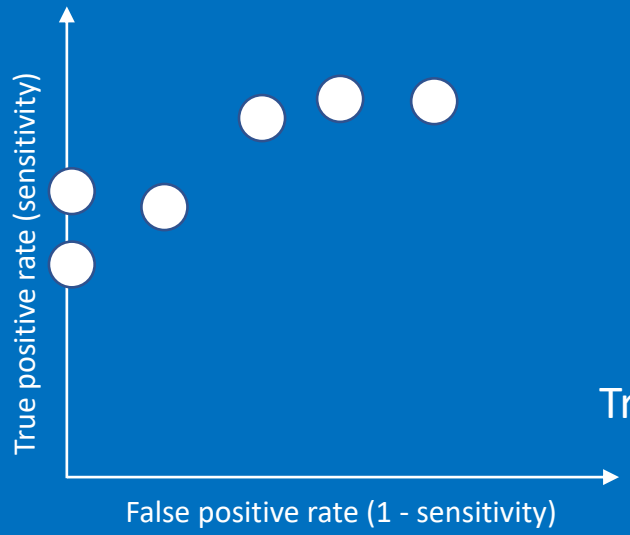
Threshold=0.2		Actual	
		Is Obese	Not Obese
Prediction	Is Obese	5	0
	Not obese	1	3

.....

True positive rate tells you what proportion of obese samples that are correctly classified

False positive rate tells you what proportion of obese samples that are NOT correctly classified

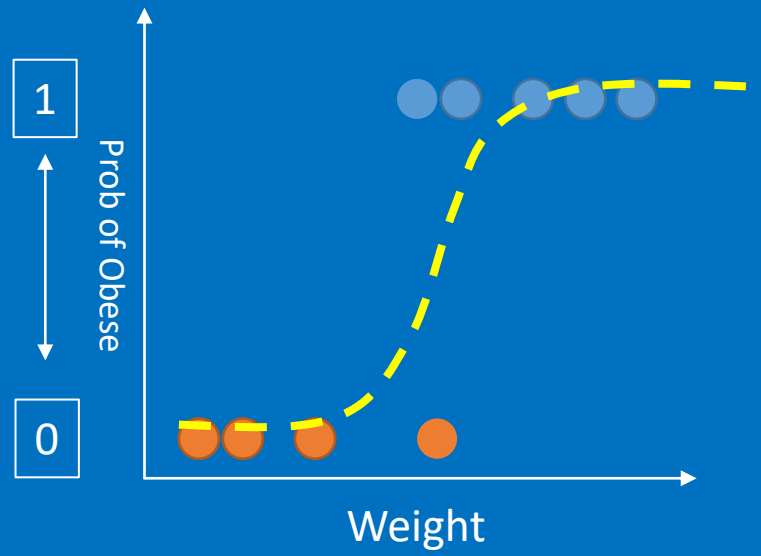
ROC score is designed to simplify the visualization of such huge number of matrices …



So by plotting all the metrics on the graph, we can have the left (assuming we use 5 thresholds)

True False positive rate can be calculated as

$$False\ positive\ rate = \frac{False\ positives}{False\ positives + True\ negatives}$$



We can create many confusion matrix for each selected thresholds

Threshold=0.5		Actual	
		Is Obese	Not Obese
Prediction	Is Obese	4	1
	Not obese		

Threshold=0.2		Actual	
		Is Obese	Not Obese
Prediction	Is Obese	5	0
	Not obese	1	3

.....

True positive rate tells you what proportion of obese samples that are correctly classified

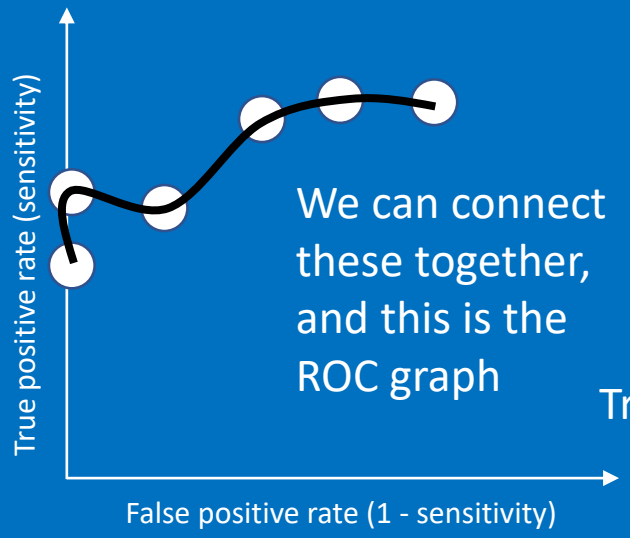
False positive rate tells you what proportion of obese samples that are NOT correctly classified

ROC score is designed to simplify the visualization of such huge number of matrices …

Let's use logistic regression, and the above data as an example

True positive rate can be calculated as

$$True\ positive\ rate = \frac{True\ positives}{True\ positives + False\ negatives}$$

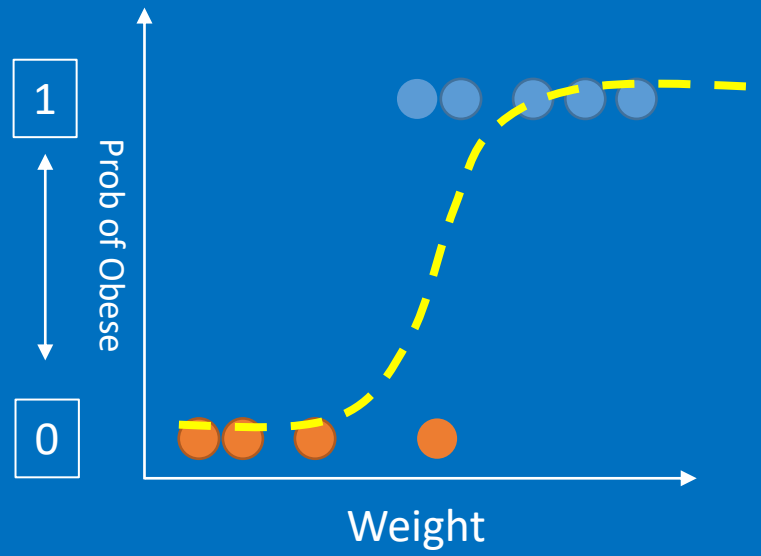


We can connect these together, and this is the ROC graph

So by plotting all the metrics on the graph, we can have the left (assuming we use 5 thresholds)

True False positive rate can be calculated as

$$False\ positive\ rate = \frac{False\ positives}{False\ positives + True\ negatives}$$



We can create many confusion matrix for each selected thresholds

Threshold=0.5		Actual	
		Is Obese	Not Obese
Prediction	Is Obese	4	1
	Not obese		

Threshold=0.2		Actual	
		Is Obese	Not Obese
Prediction	Is Obese	5	0
	Not obese	1	3

.....

True positive rate tells you what proportion of obese samples that are correctly classified

False positive rate tells you what proportion of obese samples that are NOT correctly classified

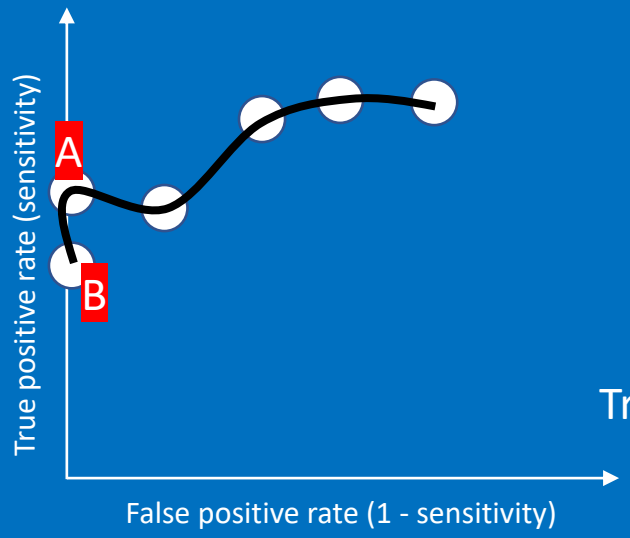
ROC score is designed to simplify the visualization of such huge number of matrices …

Let's use logistic regression, and the above data a example

From the graph, we can immediately tell: A is better than B on classification

True positive rate can be c

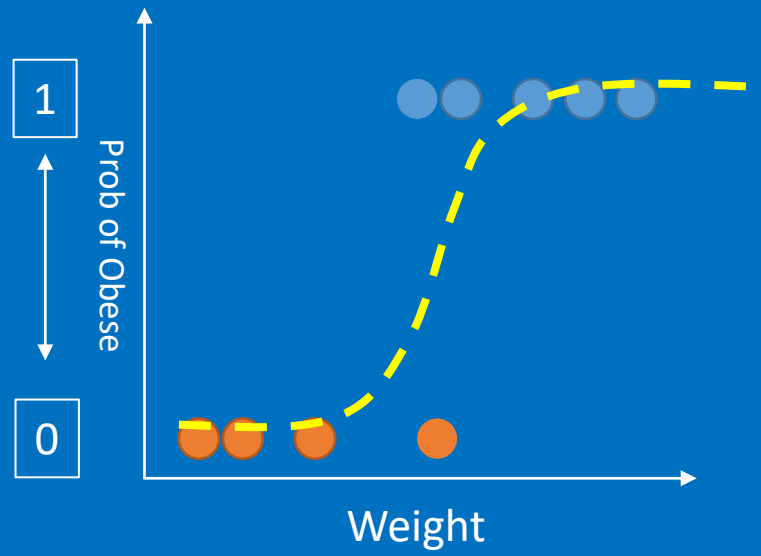
$$True\ positive\ rate = \frac{True\ positives}{True\ positives + False\ negatives}$$



So by plotting all the metrics on the graph, we can have the left (assuming we use 5 thresholds)

True False positive rate can be calculated as

$$False\ positive\ rate = \frac{False\ positives}{False\ positives + True\ negatives}$$



We can create many confusion matrix for each selected thresholds

Threshold=0.5		Actual	
		Is Obese	Not Obese
Prediction	Is Obese	4	1
	Not obese		

Threshold=0.2		Actual	
		Is Obese	Not Obese
Prediction	Is Obese	5	0
	Not obese	1	3

.....

True positive rate tells you what proportion of obese samples that are correctly classified

False positive rate tells you what proportion of obese samples that are NOT correctly classified

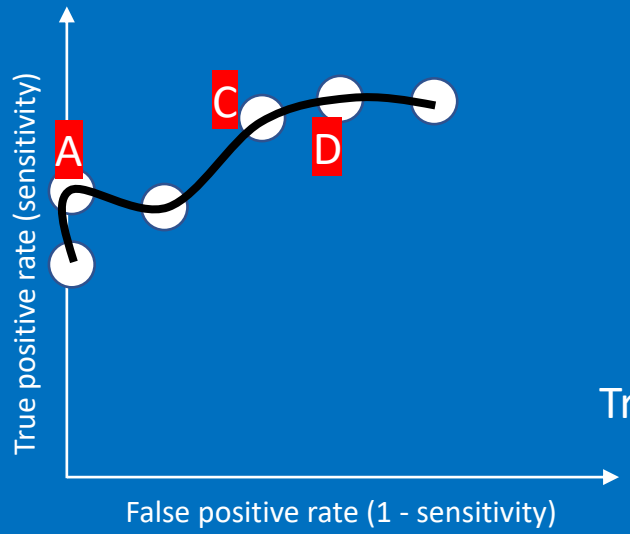
ROC score is designed to simplify the visualization of such huge number of matrices …

Let's use logistic regression, and the above data a example

True positive rate can be c

$$\text{True positive rate} = \frac{\text{True po}}{\text{True positives} +}$$

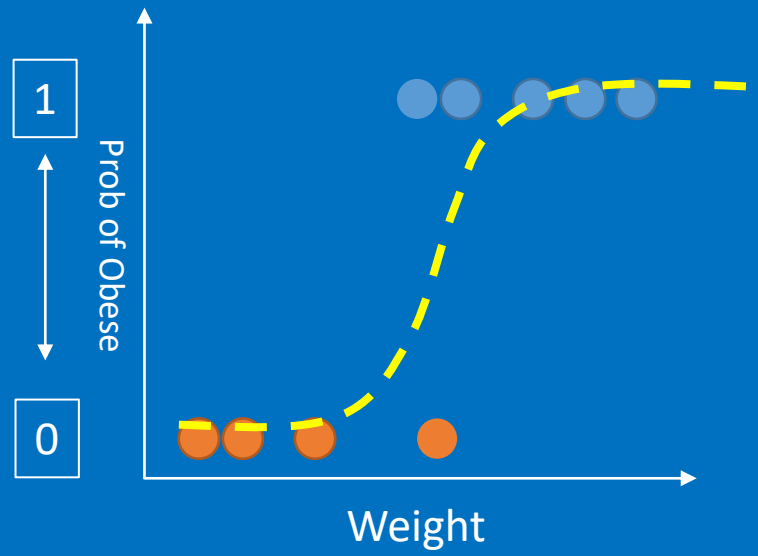
Depending on how many False positive (False alarm) I'm willing to accept, the optimal threshold is either A, C or D



So by plotting all the metrics on the graph, we can have the left (assuming we use 5 thresholds)

True False positive rate can be calculated as

$$\text{False positive rate} = \frac{\text{False positives}}{\text{False positives} + \text{True negatives}}$$



We can create many confusion matrix for each selected thresholds

Threshold=0.5		Actual	
		Is Obese	Not Obese
Prediction	Is Obese		
	Not obese		

Threshold=0.2		Actual	
		Is Obese	Not Obese
Prediction	Is Obese	5	0
	Not obese	1	3

.....

True positive rate tells you what proportion of obese samples that are correctly classified

False positive rate tells you what proportion of obese samples that are NOT correctly classified

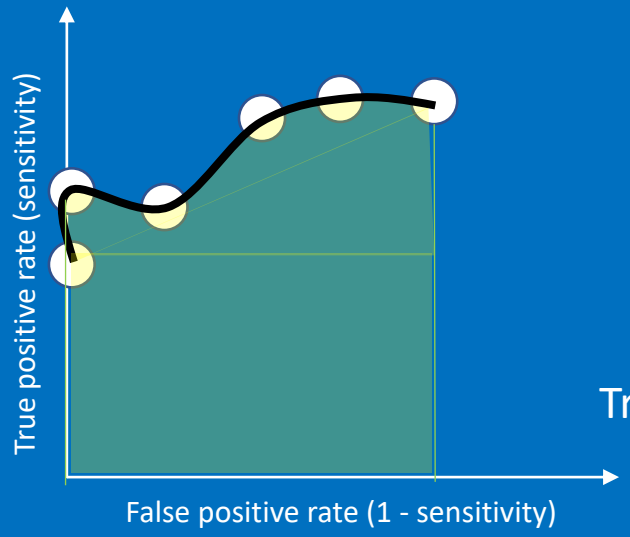
ROC score is designed to simplify the visualization of such huge number of matrices …

Let's use logistic regression, and the above data a example

True positive rate can be c

$$True\ positive\ rate = \frac{True\ po}{True\ positives +}$$

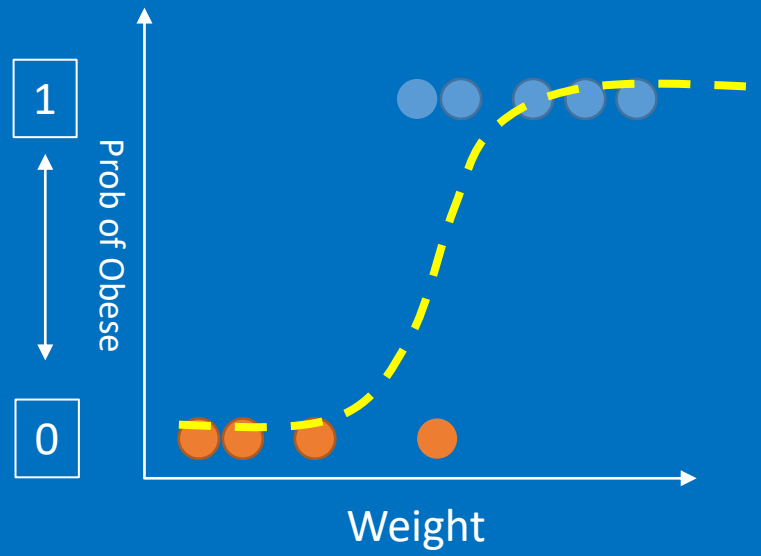
The area under the ROC is called AUC, it is used to compare different algorithm



So by plotting all the metrics on the graph, we can have the left (assuming we use 5 thresholds)

True False positive rate can be calculated as

$$False\ positive\ rate = \frac{False\ positives}{False\ positives + True\ negatives}$$



We can create many confusion matrix for each selected thresholds

Threshold=0.5		Actual	
		Is Obese	Not Obese
Prediction	Is Obese	4	1
	Not obese		

Threshold=0.2		Actual	
		Is Obese	Not Obese
Prediction	Is Obese	5	0
	Not obese	1	3

……

True positive rate tells you what proportion of obese samples that are correctly classified

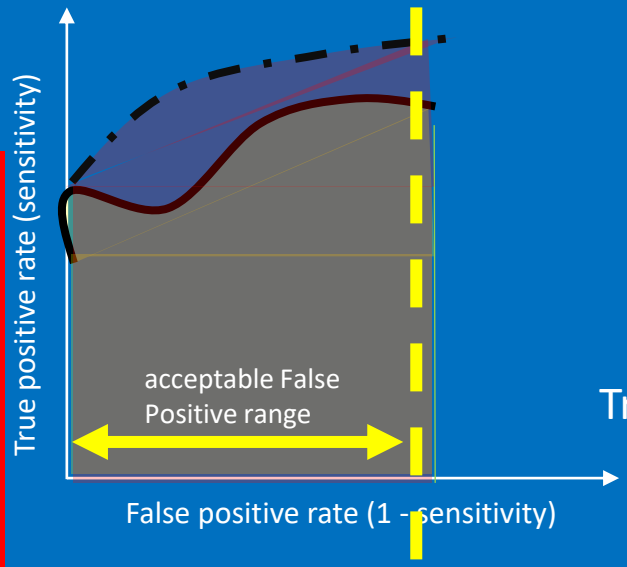
False positive rate tells you what proportion of obese samples that are NOT correctly classified

ROC score is designed to simplify the visualization of such huge number of matrices …

Let's use logistic regression, and the above data as an example

True positive

If the dot AUC represents the algorithm RF, and the solid dot AUC represents linear regression, then RF is better since the AUC area is bigger (with the acceptable False Positive range)



So by plotting all the metrics on the graph, we can have the left (assuming we use 5 thresholds)

True False positive rate can be calculated as

$$False\ positive\ rate = \frac{False\ positives}{False\ positives + True\ negatives}$$